

# World Record: HPE achieves multiple #1 performance results for AI inference benchmarks

**#1 on more than 10 MLPerf Benchmark models—HPE ProLiant Compute DL384 Gen12 and HPE ProLiant DL380a Gen11 servers**



As part of MLCommons, an AI engineering consortium, MLPerf™ Inference: Datacenter benchmarks set industry-wide standards for fairly assessing and evaluating diverse AI/ML performance across different hardware platforms.<sup>1</sup>

## HPE ProLiant Compute DL384 Gen12 Server—outstanding performance per GPU on MLPerf Inference: Datacenter benchmark



**#1**

### Stable Diffusion XL (SDXL)<sup>3</sup>

An advanced image generation model that produces high-quality, detailed images from text descriptions



**#1**

### DLRM-v2-99<sup>2,3</sup>

A deep learning recommendation model (DLRM) designed for high-accuracy prediction tasks



**#1**

### DLRM-v2-99.9<sup>2,3</sup>

An advanced DLRM optimized to achieve 99.9% accuracy in prediction tasks



**#1**

### Mixtral-8x7B<sup>2,3</sup>

An efficient AI model that outperforms larger models such as Llama 2 70B using fewer parameters

## MLPerf Inference: Datacenter v4.1 results on HPE ProLiant Compute DL384 Gen12 Server<sup>4</sup>



Best server performance with a single accelerator

Benchmark tests	Server <sup>2</sup>	Offline <sup>3</sup>
SDXL		2.31
Mixtral-8x7b	7450.72	8063.02
DLRM-v2-99	81,009.60	87,052.70
DLRM-v2-99.9	51,014.20	53,611.90

HPE ProLiant Compute DL384 Gen12 Server is an ideal solution for low-latency data center inference.

Hewlett Packard Enterprise is the first to submit performance results with the NVIDIA GH200 NVL with 144GB HBM3e memory.

## HPE ProLiant DL380a Gen11 Server top performer on 4 benchmarks



**#1**

### Image classification<sup>1,5</sup>

Resnet50 Server benchmark



**#1**

### Object detection<sup>1,6</sup>

Retinanet Server benchmark



**#1**

### Speech-to-text<sup>1,7</sup>

RNNT Server benchmark



**#1**

### Large Language Model<sup>1,8</sup>

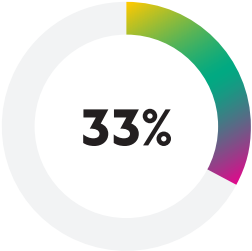
Llama 2 70B benchmark

**#1**

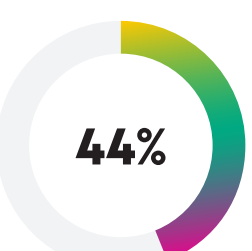
## LLM inference on HPE ProLiant DL380a Gen11 Server



Superior performance



better than the next top-performing server with 94 GB GPUs<sup>1,8</sup>



better than the next top-performing server with 80 GB GPUs<sup>1,9</sup>

<sup>1</sup> MLPerf™ Inference: Datacenter v4.1 and v4.0 as of August 28, 2024. Retrieved from [mlcommons.org/benchmarks/inference-datacenter/](https://mlcommons.org/benchmarks/inference-datacenter/). See [mlcommons.org](https://mlcommons.org) for more information. Results verified by MLCommons Association.

<sup>2</sup> Server: Scenario representing low-latency inference applications. Mixtral-8x7b: tokens per second; DLRM-v2-99 and DLRM-v2-99.9: queries per second.

<sup>3</sup> Offline: Scenario representing high-batch size inference applications. SDXL: Samples per second; Mixtral-8x7b: tokens per second; DLRM-v2-99 and DLRM-v2-99.9: samples per second.

<sup>4</sup> Based on results for NVIDIA GH200 NVL Grace Hopper Superchip with 144 GB HBM3e memory compared with all other GH200 systems (Submission ID 4.1-0072).

<sup>5</sup> MLPerf Inference: Datacenter v4.1 Resnet50 Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel® Xeon® Gold 6530 processors and four NVIDIA H100-NVL-94GB GPUs (Submission ID 4.1-0032).

<sup>6</sup> MLPerf Inference: Datacenter v4.1 Retinanet Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Gold 6530 processors and four NVIDIA H100-NVL-94GB GPUs (Submission ID 4.1-0032).

<sup>7</sup> MLPerf Inference: Datacenter v4.0 RNNT Server benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel® Xeon® Platinum 8468 processors and four NVIDIA H100-PCIe-80GB GPUs (Submission ID 4.0-0048).

<sup>8</sup> MLPerf Inference: Datacenter v4.1 Llama 2 70B benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Gold 6530 processors and four NVIDIA H100-NVL-94 GB GPUs (Submission ID 4.1-0032).

<sup>9</sup> MLPerf Inference: Datacenter v4.0 Llama 2 70B benchmark based on HPE ProLiant DL380a Gen11 Server utilizing Intel Xeon Platinum 8468 and four NVIDIA H100-PCIe-80GB GPUs (Submission ID 4.0-0048).

Visit [HPE ProLiant Compute DL384 Gen12](#)

Visit [HPE ProLiant Compute DL380a Gen11](#)